# Metadata

Gregory, Arofan; Heus, Pascal; Ryssevik, Jostein

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

**Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:**
SSG Sozialwissenschaften, USB Köln

**Empfohlene Zitierung / Suggested Citation:**
Gregory, A., Heus, P., & Ryssevik, J. (2009). *Metadata.* (RatSWD Working Paper Series, 57). Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-409531

German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

# RatSWD
*Working Paper Series*

Working Paper No. 57

# Metadata

Arofan Gregory, Pascal Heus, Jostein Ryssevik

March 2009

Federal Ministry
of Education
and Research

# Working Paper Series of the Council for Social and Economic Data (RatSWD)

---

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

*RatSWD Working Papers* are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

# Metadata

**Arofan Gregory, Pascal Heus, Jostein Ryssevik**

*Open Data Foundation* (info@opendatafoundation.org)

## Abstract

Metadata, or data about data, play a crucial rule in social sciences to ensure that high quality documentation and community knowledge are properly captured and surround the data across its entire life cycle, from the early stages of production to secondary analysis by researchers or use by policy makers and other key stakeholders. The paper provides an overview of the social sciences metadata landscape, best practices and related information technologies. It particularly focuses on two specifications - the Data Documentation Initiative (DDI) and the Statistical Data and Metadata Exchange Standard (SDMX) - seen as central to a global metadata management framework for social data and official statistics. It also highlights current directions, outlines typical integration challenges, and provides a set of high level recommendations for producers, archives, researchers and sponsors in order to foster the adoption of metadata standards and best practices in the years to come.

Keywords: social sciences, metadata, data, statistics, documentation, data quality, XML, DDI, SDMX, archive, preservation, production, access, dissemination, analysis

## I. What is metadata?

*Metadata* is a difficult term to define - it means many things to so many different audiences. If we turn to Wikipedia[1], we find: "Metadata (meta data, or sometimes metainformation) is 'data about data', of any sort in any media." While broadly true, the Wikipedia definition does not capture the real importance of metadata to those involved in social science research.

Within any domain, the term *metadata* can be more usefully defined by describing its agreed use – social sciences research has a well-developed metadata culture, which allows us to be very specific. Researchers understand what data are – the data sets which are collected, processed, analyzed and used in the conduct of research. *Metadata* is all the documentation about that data.

Even so, we are left with a definition of the term which is still incredibly broad. It is sometimes helpful to think about the different types of metadata, using common terms:

- *Structural metadata* describes the structure of data sets, whether these are tabular in nature or simply files of raw data or microdata. Which variable's value appears in which column? Which row represents which case? Are there hierarchical relationships? Etc.
- *Reference metadata (also known as "descriptive" metadata)* consists of what is often thought of as "footnote" metadata, whether this is about methodology, sampling, quality measurements, production notes, etc. This is a very broad term, which can cover a range of information, regarding everything from single data values to entire collections of data.
- *Administrative metadata* is the data which is created by the process of administering data, in its collection, production, publication, or archiving.
- *Behavioral metadata (also known as "paradata")* is information about the reaction and behavior of users when they are working with data, and respondents while data is being collected (in this case, it is paradata about a collection instrument). This can be of interest to those who act as data librarians – to help them better manage their data collections – but can also be of direct interest to researchers – what did other researchers do with the data? How did respondents react when asked a question?

1 http://en.wikipedia.org/wiki/Metadata.

It is worth noticing that metadata are for human as well as machine consumption. Whereas most of the structural metadata are there to allow software processes to read, manipulate and exchange data files, the purpose of reference and behavioral metadata is to enable human researchers to find, understand, and assess the quality of the data.

One of the criticisms of metadata as a broad discipline is that it is context-dependent, especially in terms of its use to help navigate the contents of the Internet as a whole. Indeed, there is a long-standing and on-going debate about the value of metadata. This debate – while both amusing and interesting – is not particularly relevant to those in the social sciences research domain, because very specific definitions of relevant metadata exist in the form of several standard metadata models: the Data Documentation Initiative[2] (DDI), ISO-TS 17369 Statistical Data and Metadata Exchange[3] (SDMX), Dublin Core Metadata Initiative[4] (DCMI), ISO/IEC 11179[5], the Neuchatel models for variables and classifications, and others.

The benefit of having such standards is that they allow for *direct implementation* of metadata-driven systems and management systems for metadata – and thus realization of the benefits – without having to answer questions about the precise value and meaning of metadata in its broadest sense.

## II. Metadata and technology

### *A. Historical Technology Approaches*

Metadata is a very natural part of most modern technological implementations, given the strong focus modern technology places on information. If technology depends on the *exchange and use of information* – or data – then the metadata describing that information can be very critical in the creation of systems which perform tasks in an automated way.

Early discussions about metadata were frequently concerned with describing the structure of data, whether this was the description of a simple textual format for a data file, or the structural information about a relational database schema. Other discussions were concerned more with the content of the data – that is, what sort of file is it, and what does it contain? This focus arose naturally from the ability of computers to perform computation very rapidly

2 http://ddialliance.org/
3 http://sdmx.org/
4 http://dublincore.org/
5 http://metadata-standards.org/

– the first challenge was to handle the data itself, and to perform some operation on it. Once achieved, the question was how to retain enough information about the data so that it could be *exchanged with others or used in the future*. This is where the interest in metadata came from.

It is interesting to note how little the metadata capabilities of many statistical tools have grown since the era before the Internet – even while many other types of applications assume the ability to process and understand files from other users, based on standard formats and models, statistical processing applications do not have a rich, "networked" view of the world. Many statistical tools today are reminiscent of applications dating from the 1980s – they understand enough metadata to handle specific data files, and to interpret their contents and format, or perform analytical operations, but have little ability to exchange this information with other systems or describe the context in which the data was produced.

### *B. Metadata and the Internet*

The single most important development driving the current interest in metadata is the advent of the Internet. A vast network of connected computers requires a large set of standard protocols, to allow for computers to use files from around the network. These protocols are mostly metadata.

To give a simple example: when a browser on your computer encounters a Web page, it gets a set of information from the server – metadata – which it uses to properly display that page. The Web page will probably be in HTML[6], but it might also be a Word document or a PDF file, or even a video clip. Each of these files requires a different application behavior. Thus, part of the metadata given to the browser is the MIME-type[7] of the file, which tells my computer which application to launch.

Early Internet protocols provided enough metadata to allow for human users to exchange files, but there was typically insufficient metadata for computer applications to directly perform tasks without human intervention. Because the emphasis was on people viewing files from around the network, there emerged metadata standards which supported this type of application – the best-known of these were a set of citation fields for describing any kind of resource, the Dublin Core.

---

6 http://en.wikipedia.org/wiki/HTML.
7 http://en.wikipedia.org/wiki/MIMEhttp://en.wikipedia.org/wiki/MIME.

As the Internet evolved, there has been an increasing emphasis on interactions between applications - a phenomenon termed “distributed computing”. This development pointed out that the available metadata – even with the help of standards such as the Dublin Core – was insufficient. In all of its applications, however, the Internet placed a strong emphasis on the use of remote resources without the need for explicit, human-guided integration, thus demanding a large amount of metadata, and increasingly placing importance on metadata standards.

### *C. Metadata and XML-Based Technologies*

One of the biggest developments in the growth of the Internet – and for distributed computing generally – was the advent of the eXtensible Markup Language[8] (XML), and the suite of related technologies and standards. Derived from a technology standard for marking up print documents – the Standard Generalized Markup Language[9] (SGML) – the original focus of XML was to better-describe documents of all sorts, so they could be used more effectively by applications discovering them on the Internet.

XML is a meta-language used to describe tag-sets, effectively injecting additional information into a document. Unlike HTML (which was also based on SGML), however, there was no fixed list of tags – the whole point is that documents could be designed to carry specific additional information about their contents. Thus, XML document types could be designed to carry any sort of metadata, in-line with the contents of the document.

XML is not only a language but also a collection of technologies available to perform various operations on the underlying data or metadata: XML schema, for describing document structure; XPath[10] and XQuery[11] for querying and searching XML; SOAP[12] or REST[13] to facilitate the exchange of information; and many others.

Most importantly, the above technologies are often readily available on most computers, and are free to use. The XML standards themselves are maintained by the World Wide Web

8 http://en.wikipedia.org/wiki/XML and http://www.w3.org/XML/
9 http://en.wikipedia.org/wiki/SGML.
10 http://en.wikipedia.org/wiki/XPath.
11 http://en.wikipedia.org/wiki/XQuery.
12 http://en.wikipedia.org/wiki/SOAP.
13 http://en.wikipedia.org/wiki/Representational_State_Transfer.

Consortium[14] and publicly available. This implies that XML not only provides a common language and facilitates metadata management but is also easy to adopt as a technology. While XML does not preclude the existence of legacy metadata management systems, it has shifted the way we model the information structure and expose the metadata to the outside world. Harmonized models have emerged in various field of expertise, including the social sciences.

The Dublin Core was quickly realized in an XML format, and other standards also used the new format, notably the Data Documentation Initiative (DDI – see below). At first, these standards were very much designed with human users in mind, but those involved with solving problems related to distributed computing realized that XML was a very powerful tool as well.

These developments lead to a set of Web services[15] standards (SOAP, WSDL, etc.) as well as a new type of service-oriented architecture[16] (SOA). The development of Web-services technology and service-oriented architectures continued the demand within applications for exactly-defined metadata exchanged using standard protocols. Some of the later standards such as SDMX – and later versions of existing standards (such as DDI version 3.0) are designed to leverage these developments.

Today, we have a powerful set of technology tools and metadata models which are directly relevant to the applications used by the social sciences researcher. While not all of the statistical software packages have leveraged these developments, increasingly we are seeing uses of these new metadata-rich technologies, to provide functionality to researchers and those who support them which were not possible with earlier generations of technology.

## III. Metadata and social-science

### *A. Why metadata?*

In social sciences, the quality of the data has a direct impact on the soundness of policies or validity of the research outputs. Data quality is typically measured using criteria such as accessibility, coherence, relevance, timeliness, integrity, consistency, coherence, amongst

14 http://www.w3.org/

15 http://en.wikipedia.org/wiki/Web_service.

16 http://en.wikipedia.org/wiki/Service-oriented_architecture.

others. These indicators are generally accepted as a good measure of the overall usefulness of the data. Meeting these criteria not only means making data available but also requires comprehensive documentation of the data structures, production processes, statistical methodologies, data sources, context, and many other aspects. This is necessary not only for usability purposes but also for discovery, accessibility, preservation and information exchange purposes.

In social sciences, metadata is therefore essential for several reasons:

- It is a requirement to ensure that sufficient information is available to the users in order to properly *understand* and *use* the data. Without relevant documentation, researcher would be unable to properly interpret the meaning of the data. Lack of information also puts extra burden on the data provider who needs to respond to user's queries.
- It is necessary to facilitate the *discovery* and *access* of the data by its intended consumers. The best data in the world is useless if no one is aware of its existence.
- It supports the long term *preservation* of data by ensuring that relevant information remains with the data for its future use or for conversion into new archival formats.
- Common metadata language and structures are also essentials to support the *exchange* of information between agencies and/or individuals.

In general, better documentation makes for more useful data, and ultimately better research. The usability of data is intricately tied up with issues about how completely it is documented – rich metadata about a data set allows for easier access and use of the data. Researchers want better data, and one way to help improve data quality is to provide better documentation.

### *B. Metadata and the data lifecycle*

The data lifecycle in social science is quite complex as the data flowing from the survey respondents or administrative systems to the researchers and policy makers goes through several stages and transformation processes involving many different actors. Furthermore, secondary or derivative data and research findings often themselves become data sources for others.

Any description of the purpose of metadata within the data lifecycle should start with an analysis of the requirements of the users:

- The majority of data users have not been engaged in the creation of the data they are

using.

- Data will frequently be used for other research purposes than intended by the creators (secondary analysis).
- Data will frequently be used many years after they are created.
- Data users are often comparing and combining data from a broad range of sources (across time and space).

The common denominator of the four characteristics is an emphasis on the relative distance between the end users of a statistical material and the production process. Whereas the creators and primary users of statistics might possess “undocumented” and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal metadata that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. Without human language descriptions of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

Ideally, data providers should abide by Gary King’s replication standard[17] that holds that “sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author.". Note that from this perspective, researchers are as much as producers defined as “data providers” and should therefore abide by the same documentation principles.

Metadata has however traditionally not been the focus of data producing agencies and the responsibility of documenting data was often in the past left to the data archive, data librarians or research data centers. Such “after-the-fact” effort requires substantial resources and typically leads to a considerable amount of information loss and sparsely documented data.

This mindset has changed in recent years, and considerable efforts are now being made by data producers and archives to improve the overall quality of metadata. The idea is also being extended to the researchers or end users whose contribution to metadata is often non-existent.

17 "Replication, Replication", Gary King, PS: Political Science and Politics, Vol. XXVIII, No. 3 (September, 1995), 443-499 and "A Revised Proposal, Proposal," Vol. XXVIII, No. 3 (September, 1995), 443-499. (Article: PDF). See also http://gking.harvard.edu/projects/repl.shtml.

Collecting inputs from the users themselves should lead to a better understanding of the data usage, reduce the duplication of efforts and promote the sharing of knowledge. This shift from a centralized maintenance of metadata by the archive to a distributed approach where many entities contribute to the knowledge seems only natural: it is better and easier to capture information about an event at the time of its occurrence rather than after the fact.

There is another view of the data lifecycle which is not so much concerned with the collection and production of data for research as it is the aggregation and harmonization of data. This view can be termed the *information chain,* because it describes the flow of data from its original micro-level source(s) through the various aggregation and harmonization processes, as the data flows upward from it source through the hierarchy of primary and secondary users. Data collected through surveys or from administrative sources at a regional level might be aggregated at a national level, combined with other sources, and then further aggregated at the international level.

This view of the data lifecycle also places importance on the distance between those collecting the original data, and its eventual use at a higher level of aggregation. Without sufficient documentation about the aggregation and harmonization processes, it is difficult for end-users to fully understand the aggregates they are using.

The main goal of capturing metadata at each stage of the lifecycle is to persist it throughout both a single cycle from collection to publication (and hence to archiving), but also to capture each secondary use of the data, so that any data set will be accompanied by as complete a set of documentation as possible. Information captured as it happens is of higher quality and completeness, which directly benefits the user of the data.

Other less-obvious benefits to having a persisted set of metadata accompanying the data through the lifecycle also exist, however – good metadata can be used to help drive the processing of the data as it flows through the lifecycle; and collections of well-documented data become available for comparison with other, similar data sets. Good information about the content and processing of a collection of data can provide valuable information to those who want to re-purpose or manage data within that collection. Thus, the beneficiaries of good metadata, captured as the data is collected, processed, and published include not only the researcher, but also the secondary user, the archivist, and the data producer.

Very often, good metadata can form the basis of code generation, whether that code is running inside a statistical package, or is used for some other purpose (such as automatic generation of forms for data collection). It can also be used for the automated production of documentation or publication that can be customized to the end user needs. Although not immediately apparent, the benefit of having good metadata is that the systems which support the researcher, data producer, and archivist can all be made much more efficient, and produce more better data.

### *C. Standard metadata models*

The recent emphasis on the data lifecycle, and capturing metadata from the beginning, has driven the development of two standard models, each designed around one of the data lifecycle views described above. The Data Documentation Initiative (DDI) is, in its most recent version, based around a lifecycle model which describes the collection and sourcing of data, through the stages of publication, archiving, and secondary use. ISO TS-17369, the Statistical Data and Metadata Exchange standard (SDMX) is based on a view of the information chain, with a stronger focus on aggregate data products. These standards – along with a number of others in various relevant areas – create a common view of how metadata within the social sciences domain can be described and exchanged, to facilitate the flow of metadata alongside the relevant data sets.

## IV. The Data Documentation Initiative

### *A. DDI – early history*

The Data Documentation Initiative[18] (DDI) is an international program to produce a metadata specification for the description of social science data resources. The program was initiated in 1994 by the Inter-University Consortium for Political and Social Research (ICPSR). Contributors to the efforts come mostly from social science data archives and libraries in USA, Canada and Europe.

The original aim of the DDI was to replace the widely used OSIRIS codebook specification with a more modern and Web-aware specification that could be used to structure the description of the content of social science data archives. The first preliminary version came

18 http://www.ddialliance.org.

in the form of an SGML Document Type Definition[19] (DTD), which in 1997 was converted to an XML DTD. The migration to XML happened just a few months after the W3C released the first working draft of the XML specification. The DDI was consequently one of the very first major metadata initiatives using the new framework. Several data archives started to use the DDI to describe their data collections and software was developed to support its use. However, it soon became clear that the first versions of the DDI had several severe limitations:

1. A pure “bottom-up” approach

The DDI specification was developed to describe concrete files or products coming out of the statistical production process. Given it roots in social science data archiving this is quite natural. The information objects of the data archives are finalized products that have cut the lifeline to the various production processes and put into the hands of the users.

As a consequence there was a one-to one relationship between a DDI instance and the physical data it was meant to describe. The DDI was tied to the dataset and there were no methods to describe abstract statistical concepts that might be represented in more than one concrete study. It was therefore impossible to reference identical variables across datasets, and even series of survey instances where the majority of variables are identical from wave to wave had to be described instance by instance.

2. Modularity

The first versions of DDI had its roots in a “book” metaphor. It was seen as a digital equivalent of a paper document – the well-established codebook or data dictionary. The specification was not built according to a modular architecture that allowed information and application providers to select bits and pieces and “snap” them together on a more freely basis.

3. Extensibility

Another critical limitation was the lack of a proper extensibility mechanism. Within the confines of an XML DTD there are no ways to add local extensions without compromising the interoperability of the core specification. You either accept the specification as it is without any additions or you break it. For a big and complex specification like the DDI this is a major problem that can easily hurt the adoption process. Without a mechanism that allows extensions to be made without breaking the standard the chances are high that application

19 http://en.wikipedia.org/wiki/Document_Type_Definition.

providers might sacrifice interoperability for local efficiency and relevance.

Despite these limitations, the DDI answered the fundamental needs of data archives for documenting survey datasets and has been widely adopted by agencies around the world.

*B. DDI version 3.0*

Version 3.0 of the DDI was released in April of 2008, representing a major revision to the standard which solved the problems found with earlier versions as described above. Based on a survey lifecycle model, it is designed to describe groups and series of studies, to define degrees of comparison within and across studies and to allow for re-use of metadata where appropriate. It uses a modular approach, with modules which are related to each step of the data lifecycle. Different types of metadata are organized into packages relating to their contents. All the metadata about a survey instrument, for example, is found in the "data collection" module, represented by an XML namespace.

DDI 3.0 represents an approach to the metadata which is more in line with the capabilities of modern information technology: it is relational in nature, rather than document-centric, so that metadata can be easily referenced and reused. This is important, because modern web-services technology leverages the idea of distributed computing. The DDI 3.0 is designed explicitly to support the concept of having a collection of metadata be distributed and re-used by reference.

The combination of the lifecycle approach, a modular design, and metadata reusability has transformed the specification from a product intended for archiving datasets by a single agency into a highly flexible standard that can be used by all actors of the survey lifecycle for different purposes. Expected uses of DDI 3.0 include study design and survey instrumentation, questionnaire generation, support for data collection and processing operations, capturing data aggregation or recoding, manage question or concept banks, data discovery, research project, data comparability, metadata mining, and likely several other purposes that cannot be foreseen. For each case, a subset of the specification is used for the specific purpose or to provide a customized view of the information. The strength of DDI 3 is that it maintains a common language and metadata consistency across the lifecycle stages and amongst contributors.

The new version has also been designed to work with standards such as SDMX, ISO 11179, Dublin Core and others which ensure that the metadata can be connected to other domains or stages of the lifecycle. It takes into account backward compatibility with previous versions of DDI to ensure that current users can continue to use their existing framework or metadata.

Overall, DDI 3.0 has broaden the scope of the specification and made the standard attractive to a broader range of users across the entire survey lifecycle, from data producers to researchers.

*C. Adoption of the DDI*

In its early stage of existence, the DDI specification was primarily used by the data archive community in North America and Europe. With only a handful of tools available, the first DDI users relied on proprietary solutions to manage their metadata or even compiled the metadata by hand! The advent of the Nesstar[20] software played a key role in the adoption and success of the DDI as the only production-grade solution. In 2006, the International Household Survey Network integrated the Nesstar Publisher as one of the components of its Microdata Management Toolkit,[21] a set of tools targeted towards national statistical agencies in developing countries for the preservation and dissemination of survey microdata. Supported by the PARIS21 / World Bank Accelerated Data Program,[22] the toolkit has met great success and is now in use in dozens of countries across Africa, Middle East, Latin America and Asia. DDI is now a truly global specification.

With the publication of DDI version 3.0, the DDI Alliance has broadened the potential user based of the specification to all agencies and individuals involved in the survey life cycle. While no official implementation of 3.0 is currently in use, several organizations (primarily producers and research data center) have expressed their interest in adopting it or are already in the initial stage of implementation. The availability of generic tools will play a major role in the success of 3.0 but, once this initial hurdle will be passed, a large uptake of the new version is expected.

20 http://www.nesstar.com.
21 http://www.surveynetwork.org/toolkit.
22 http://www.surveynetworl.org/adp.

## V. The Statistical Data and Metadata Exchange

In 2001, seven international and supra-national organizations organized the Statistical Data and Metadata Exchange[23] Initiative: the Bank for International Settlements (BIS), the Organization for Economic Cooperation and Development (OECD), the European Central Bank (ECB), Eurostat, the World Bank, the International Monetary Fund (IMF), and the United Nations Statistical Division (UNSD). The initiative was formed to examine how new technologies could be used to better support the reporting and dissemination of aggregate statistics, which all of these organizations use to support policy and development activities.

In 2005, the first version of the SDMX technical standards (that is, technology standards) became an ISO Technical Specification, ISO TS-17369. They provided an information model and XML formats for all types of aggregate data and related structural metadata, along with guidelines about how web services should be supported. There is also a legacy format in UN/EDIFACT syntax, formerly known as GESMES/TS (but now SDMX-EDI) which is still supported under the SDMX model.

Having standard XML formats for data and structural metadata made the process of exchanging data more efficient, because the data would be predictable, and would be accompanied by rich metadata. SDMX has been implemented by many additional international organizations, and national-level institutions such as central banks and statistical offices. Adoption is global.

In 2008, the SDMX Initiative released two other sets of products which are important: a second and much-expanded version of the technical specifications SDMX 2.0 (now being submitted to ISO for acceptance as an International Standard) and a set of Content-Oriented Guidelines, which recommend how various statistical concepts in broad use can be defined, named, represented, and used.

In addition to support for aggregate data sets and related structural metadata, the version 2.0 of the technical specifications provide support for all types of reference metadata, including an ability to mimic the contents of other related standards for the purposes of cross-walking. There is also a standard for providing registry services, a feature of web-services architecture

23 http://www.sdmx.org.

which allows for the easy location of data and metadata resources around a distributed network.

It is significant to note that both SDMX and DDI were designed to be aligned and to work well with other related standards – SDMX was designed with a knowledge of DDI (version 3.0 and earlier versions), and vice-versa. An effort was made to make sure that these standards are complementary, rather than competitive.

## VI. Other specifications

There are several other standards which are of interest to the social sciences researcher. These will be given a brief mention here, and the list provided is not exhaustive.

- *ISO/IEC 11179:* This standard provides a model for understanding what it terms "data elements" which are as applicable to metadata as they are to data. The model provided gives a standard way of defining terms, the concepts they represent, the value domains which they encompass, and how those value domains are represented. Additionally, a model for lifecycle management is provided. Ultimately, this is a powerful model for defining the semantics of different terms and concepts used with social sciences data.
- *ISO 19115:* This standard provides a model for defining geographies, and is used by many other systems which care about geography, maps, etc. This model is embedded in DDI, for example, but is widely used.
- *Dublin Core:* Dublin core provides a set of fields for providing the citations of resources, and has a core set and an extension mechanism, expressed in XML.
- *METS:* This is a standard from the world of digital archives, which provides for the packaging of a set of related objects (e.g., a Web page and the image files it references). It allows for other standard metadata formats to be embedded in it (DDI is one example of this).
- *PREMIS:* This is an XML format for expressing metadata about the archival lifecycle, and is meant to be used in combination with the OAI archival reference model.

Given the many stages data and metadata goes through in social science and the different perspectives taken by the various actors, it is clear that a single metadata specification cannot be used to cover the entire life cycle. Using The DDI and SDMX as core standards and extending their functionalities through combination with the other standards mentioned above offers data producers, librarians, researchers and other consumers a robust set of tools for the

management of data and metadata across the entire lifecycle. The often non-trivial job of mapping these standards correctly to one another is being undertaken in forums such as UN/ECE's METIS[24] conference and elsewhere.

One example of this is the use of DDI to document micro-level data sources, with resulting aggregates described using SDMX. Each standard is best suited to a different set of processes – having them well-aligned, and mapped, allows for the combined use of the standards in an efficient and consistent manner.

## VII. Metadata in Germany

There has been much involvement from some German organizations in the development and use of metadata standards, and today, Germany is one of the leading countries in terms of adoption of the standards described in this paper. Our impression is that the more recent interest in DDI and other standards such as SDMX is being driven at least partly by legislative changes regarding the exchange of data between state-sponsored institutes, but we are not familiar enough with German law to make any definite pronouncement. Certainly, German involvement in metadata standards has a long history.

The involvement of Germany in the creation of metadata standards focuses mostly on DDI – some German institutes such as GESIS were very involved in both the development of past versions of the standards, and also in their implementation. The German micro-census is a good example of how DDI was – and continues to be – used for data documentation, but there are many others.

More recently, some of the other German institutes involved with social sciences and economics have started using DDI, and participating actively in the DDI community. Most notably within research data centers (RDCs), where application must be made to access confidential data, there has been increasing uptake of and interest in the use of DDI 3.0. This reflects an international trend, but Germany is one of the most active countries for RDC developments in the use of DDI. At the IASSIST 2008[25] conference at Stanford University, the Institute for Employment Research (IAB)[26] presented a prototype for using the DDI 3.0

24 http://www.unece.org/stats/archive/04.01d.e.htm.
25 http://iassist08.stanford.edu.
26 http://www.iab.de/

metadata model as the basis of a documentation system which will serve both the RDC and the internal research departments. At the International Data Service Center of the Institute for the Study of Labor (IZA)[27] in Bonn, DDI 2.1 is used as the standard metadata model, and, in future, DDI 3.0 will be used.

One reason for the leadership role exhibited by Germany within the social sciences metadata community is the hosting of DDI-related events for the past two years at Schloss Daghstul, a Liebnitz Institute focused on computer science conferences. Organized by GESIS, with some co-sponsors, there have been focused seminars given to help provide an in-depth understanding of DDI 3.0, and also other DDI-related meetings on related themes (in 2008, the topic was DDI 3.0 best practices). These have taken place in the fall of 2007 and 2008, and it appears that these events will become an annual feature of the DDI community calendar. They have attracted attendees from all over the world.

In 2009, the first European DDI User's Group meeting will be hosted by IZA, which has also played a significant role in organizing the group. Thus, it can be seen that German institutes have had a significant role in the development and use of DDI, and this role appears to be growing with the advent of DDI 3.0.

SDMX has also been supported within Germany. DESTATIS in Wiesbaden was an early participant in the SDMX Open Data Interchange (SODI) project run by Eurostat, along with a small number of other European national statistical organizations. The European Central Bank in Frankfurt – although not a German organization as such, but a European one – is one of the sponsors of SDMX (along with the BIS, the IMF, the OECD, Eurostat, the World Bank, and the UN Statistical Division), and was also a major user of the standard on which SDMX was based, GESMES/TS.

Increasingly, there is a growing interest in the exchange of research data and statistical data both within countries and across national borders. Metadata standards such as DDI and SDMX are a critical ingredient to facilitating these exchanges. Germany has emerged as one of the more forward-looking countries in this respect.

---

27 http://idsc.iza.org/

## VIII. Directions, Challenges and Recommendations

The availability of high quality metadata promises to drive many positive changes within social sciences in the near future. Better metadata allows for better use of technology, which can fundamentally impact what is possible for researchers: (1) data which is better documented, easier to find and use, and is of greater consistency and quality; (2) heightened visibility for researchers' findings, and the ability to replicate and validate those findings using the actual data and processes; (3) new techniques for identifying comparable data sets, and an increased level of granularity in working with data from multiple sources; (4) improved tools for data management, to assist data producers, librarians, and archives; (5) and the establishment of virtual research communities.

It is worth noting that important components of the technology suite needed to realize these benefits are *web-services*[28] based architectures and *registries.*[29] The first is the industry standard technology essential for allowing applications to effectively communicate with each other and exchange information. The second implements public catalogues for applications within a domain to facilitate searching and locating data and metadata resources wherever they are located on the Internet or network. This combination is essential to support the establishment of dynamic portals and federated spaces that provides users with a virtual view of the statistical information and effective mechanisms for timely publication of data, documents and research outputs. It also unlocks powerful features such as notifications services (whereby the information flows automatically towards its intended users, not the other way around), comparability and harmonization, researcher feedback, and community driven knowledge spaces.

Another significant emerging idea is the concept of *enhanced publications*, which combine research findings, data, and metadata as a single package, providing support for the replication standard within social sciences. Given a collection of such publications, it becomes possible to maintain linkages between primary and secondary datasets and publications, providing for richer comparison and broader knowledge. Well packaged information also allows for the use of data at the level of the variable, rather than just the monolithic dataset, supporting more granular comparison and exploration by topic.

28 http://en.wikipedia.org/wiki/Web_service.
29 http://en.wikipedia.org/wiki/Metadata_registry.

These benefits will not be realized without meeting some significant challenges, however. These can be broken out into three categories: (1) Tools; (2) Metadata quality; and (3) Practice. Most agencies or individuals will likely confront issues in each of these areas, but it is important to know that they do not need to do so in isolation. Organizations such as the Open Data Foundation, the DDI Alliance, the International Household Survey Network and others are working towards bringing together users for the purpose of sharing resources and expertise to jointly address metadata challenges.

**Tools:** An XML specification by itself is not something that can be used out of the box. It requires software to allow for the capture, storage, publication and exchange of the metadata. Building such products can be an expensive effort and this problem was recognized by the DDI and SDMX sponsors. To address the issue, several initiatives are ongoing for the development of open source solutions to facilitate the use and adoption of DDI and SDMX. The DDI Foundation Tools Program[30] aims at the implementation of a DDI 3.0 core framework and utilities for implementers as well as the production of a generic DDI 3.0 editor. The Open Data Foundation is working with its partners to release a free SDMX browser tool and provides a source code repository to anyone interested in developing open source software for social-science metadata management. The International Household Survey Network has also developed a DDI 2 based Microdata Management Toolkit targeted for statistical agencies in developing countries.

We therefore recommend anyone interested in adopting metadata standard to check with relevant organizations regarding tools availability or even contribute to the joint development efforts.

**Metadata quality:** Having tools available does not mean that the metadata will be sound and reliable. In the end, it is the content that counts and compiling high quality comprehensive metadata also requires good techniques, guidelines and a significant amount of discipline. While some of the work can be automated or semi-automated using software utilities, it is often necessary to compile information by hand and chase down metadata to find the missing piece of knowledge or document. This is particularly true when the metadata is captured after the fact or back logging. This implies that human error and missing information are a factor.

30 http://tools.ddialliance.org.

Quality assurance is therefore very important aspect of metadata management and any organizations adopting standards should thoroughly document these processes. As a general rule, metadata should be treated as an official publication and should therefore follow the same institutional rules.

Harmonization of practices across organizations also plays a major role when the metadata leaves the institution and is shared with users or other partners. If the same metadata elements are documented using different principles, it will no longer be coherent which can confuse users, impact comparability and reduce system interoperability.

Agencies such as the DDI Alliance, the International Household Survey Network or SDMX sponsors are producing generic guidelines and best practices for the preparation of metadata. They also work closely with metadata producers toward the harmonization of metadata elements. When looking into metadata quality assurance issues, we therefore suggest consulting existing web site and the literature for references or joining existing initiatives. We also recommend for agencies working in smaller communities to actively collaborate towards metadata harmonization.

**Practice:** Adopting new standards and technologies implies a change in the way the organizations and individuals have been operating. While the benefits of a sound metadata management framework are extensive, this inevitably meets some resistance and requires a certain amount of resources to foster acceptance. Just because the tools and guidelines exist to help realize the benefits does not mean that people will use them. Researchers, especially, are often reluctant to recognize that new techniques and discipline are necessary. Awareness, training and integration are all adoptions issues facing researchers, archives and data providers. Highlighting the benefits and providing incentives will be necessary to lead to successful integration.

Given the strong interest displayed by data providers towards metadata standards, we do anticipate the rate of adoption for DDI and SDMX to continue to pick up strongly in the coming years. A key to this success will be the availability of generic software tools. Sponsors and community driven open source initiatives are expected to contribute a wide range of generic products for the management, publication and sharing of metadata that will foster the standard adoption. These initial efforts will likely start to show significant results in 2009-

2010. In the meantime, statistical agencies and research data centers with strong internal IT capacity will likely in parallel design their own tools to manage metadata. As the potential market grows in size, it is also possible that statistical packages or other commercial vendors will begin to provide solutions as well.

While the metadata will continue at first to primarily emerge from data archives, the uptake amongst producers should raise which will improve the overall quality as the information is captured closer to its source. Researchers will also likely begin to contribute to the metadata knowledge. Such end user adoption may be slow at first but incentives and benefits should quickly out weight the resistance to change or contribute and we should see an increase in user based metadata. This overall will foster the existence of shared knowledge space through metadata and bridge the communication gap that often exists between user and producer.

Given that many actors will now be contributing to the metadata, best practices and harmonization will play a crucial role in the overall quality and consistency of the information. Lead by sponsors and large statistical agencies, national and international initiatives will likely emerge to draft metadata management guidelines and work towards the harmonization of common metadata elements. This will not only lead to improved metadata but will also foster better and more comparable data.

As more and more standard metadata is being produced, the need for exchange, sharing and publication will quickly increase. As end users prefer to have single point of entry, national, regional and international catalogs or registries will grow in importance. This aggregation of information will support the development of large collection of information that could potentially support complex searches and metadata mining operations. Note that such registries do not store the actual data. They act as lookup point that are used to retrieve the location where the information actually is (just like a phone or address book).

In order to foster broad adoption of metadata and related best practices in social sciences, we recommend to:

1. Promote the importance of high quality data documentation and its capture using metadata standards
2. Familiarize producers, archives and researchers with metadata standards, related best practices and technologies

3. Support the development of standards based tools, preferably under an open source license and aligned on community recommendations
4. Do not undertake metadata adoption activities in isolation. Instead join and sponsor community or federated driven initiatives
5. For data and metadata managers and providers, support the establishment of an industry standard web service oriented and registry based IT infrastructure to facilitate the management, exchange, reuse and harmonization of metadata and data.
6. Integrate metadata capture at all stage of the life cycle. Document events as they happen, not after the facts.
7. Leverage on the availability of metadata to automate the production of documentation or generation of statistical scripts to reduce the overall production costs, increase quality and deliver user customized products.
8. Support the establishment of virtual research and collaborative spaces to allow for user driven metadata, foster community knowledge capture.

Overall, the future of social-science metadata looks very bright. The availability of robust standards combined with modern technologies has laid the foundation of a global harmonized framework for the management of social science data and documentation. Just like the Internet has revolutionized and connected our world, social-science metadata has the potential to open new possibilities for producers, archives and users.